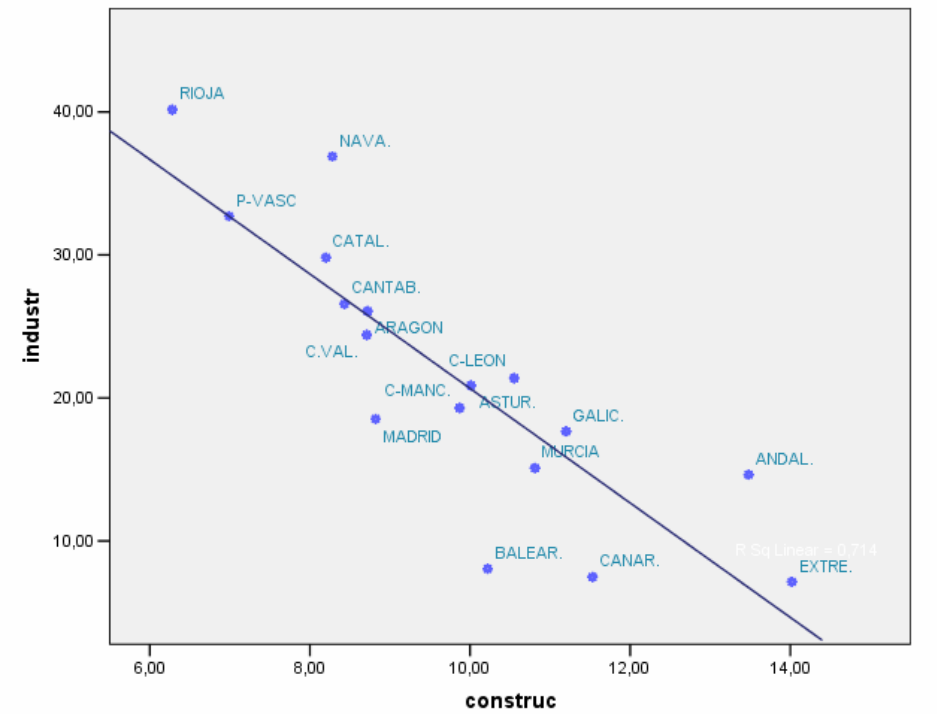
Regresión Lineal Simple



J. M. Rojo Abuín
Instituto de Economía y Geografía
Madrid, Febrero de 2007

Índice

I. INTRODUCCIÓN

Históricamente el nombre de modelos de regresión se debe a los estudios de Galton en biología. Galton, al estudiar la relación entre las estaturas de los hijos (Y) con la de sus padres (X), se dio cuenta que los hijos de padres altos son más altos que la media pero no tanto como sus padres, y los hijos de padres bajos, en general, son más bajos que la media pero no tanto como sus padres, es decir, que la altura de los hijos tiende a regresar a la media, de ahí el nombre de regresión.

En general, un modelo de regresión lineal simple consiste en estudiar la relación existente entre una variable denominada dependiente (Y) y otra variable denominada independiente o explicativa (X) a través de una recta, que toma el nombre de recta de regresión.

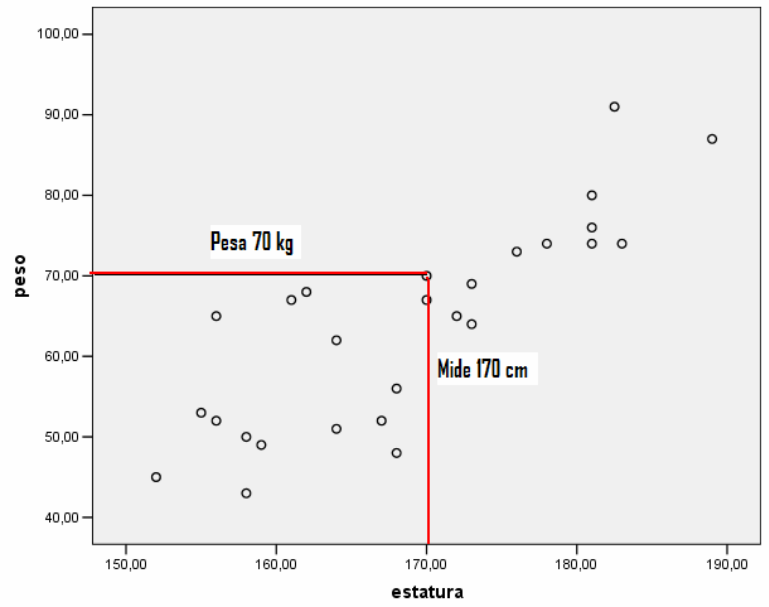
Supongamos que hemos medido, simultáneamente, el peso y la altura de una serie de personas:

Estatura en centímetros	Peso en kilogramos
X	Y
152,00	45,00
155,00	53,00
156,00	52,00
156,00	65,00
158,00	43,00
158,00	50,00
159,00	49,00
161,00	67,00
162,00	68,00
164,00	51,00
164,00	62,00

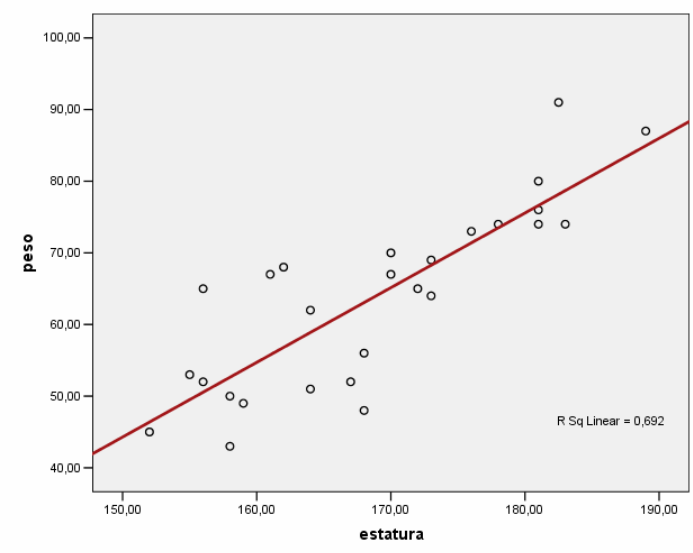
Cada fila representa los datos de un individuo y en cada columna los valores de una variable medida sobre dichos individuos.

Denominamos gráfico de dispersión a la representación grafica en un plano cartesiano de un conjunto de pares de datos (x, y).

Las observaciones pueden ser representadas en un gráfico de dispersión. En este gráfico cada individuo es un punto cuyas coordenadas son los valores de las variables.



Representando estos pares de puntos (Estatura, Peso) en un plano podemos observar la relación existente entre dicho par de variables:



En este caso se puede observar que existe una relación creciente, pues a medida que crece la estatura parece que va creciendo el peso.

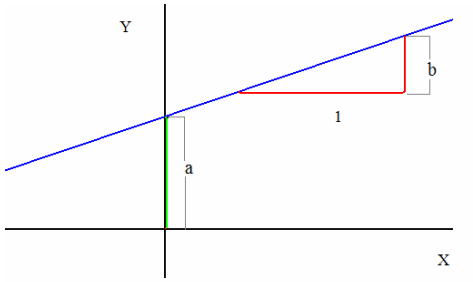
Desde un punto de vista algo simplista el análisis de regresión lineal simple, consiste en calcular una recta, denominada recta de regresión de forma que pase lo más cerca posible de todos los puntos.

El estudio de los coeficientes de la recta van a resumir la relación entre este par de variables.

La forma general de una recta es la siguiente:

$$Y = a + b * X$$

Donde **a** es la constante y **b** es la tangente; estos parámetros tienen la siguiente interpretación grafica:



Los coeficientes de la recta tienen el siguiente significado:

- **a** es el valor de Y cuando x es igual a cero.
- **b** es la pendiente o inclinación de la recta.

Siguiendo con el ejemplo anterior, la recta de regresión que deseamos calcular es:

$$Peso = a + b * Estatura$$

Y los coeficientes tendrán el siguiente significado:

- a es el valor que toma la recta de regresión cuando la estatura vale 0.
- b es el incremento en el peso de una persona por un incremento de un centímetro en su estatura.

Nota

En general, el coeficiente **a** suele ser de poco interés, pue,s en general, no estaremos interesados en el valor pronosticado de la variable dependiente cuando la variable explicativa vale 0. A pesar de todo, este coeficiente **no debe ser eliminado del modelo,** pues, en caso contrario, estamos obligando a que la recta de regresión pase por el origen de coordenadas y quitando un grado de libertad del modelo.

Es importante observar que el coeficiente **b** va a tener unidades de medida, en este caso kilos/centímetro, por lo tanto es importante especificar las unidades de medida de las variables.

II. HIPÓTESIS

Antes de realizar un análisis de regresión lineal hay que tener en cuenta qué cualidades deben o deberían de tener los datos que vamos a analizar.

En primer lugar vamos a considerar que los datos han sido generados, o son una muestra aleatoria de una población en que la relación existente entre las variables es lineal, o bien que en el intervalo muestreado se comporta de manera lineal.

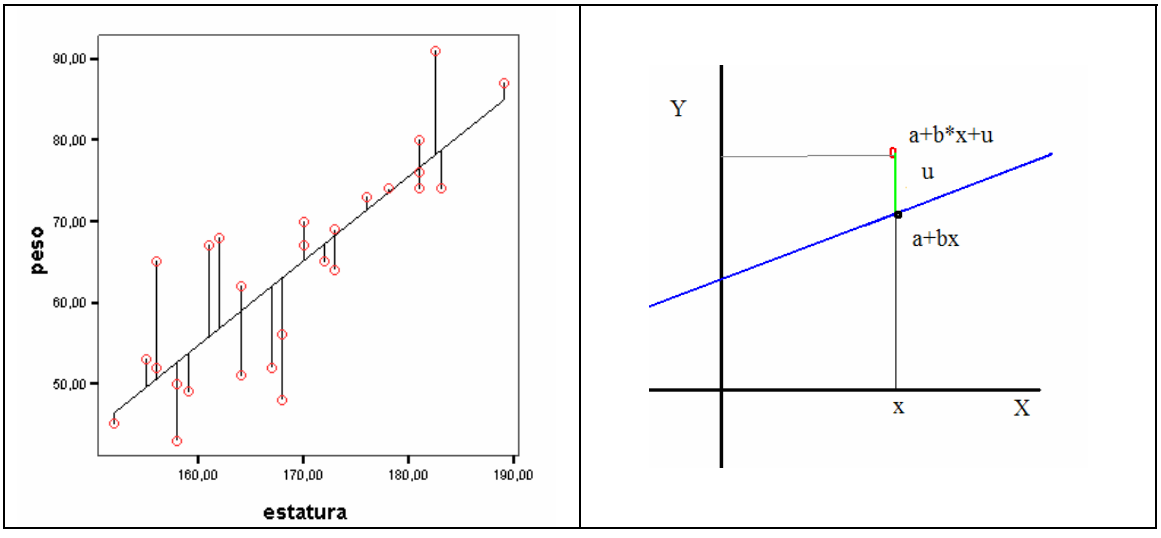
El modelo de regresión lineal simple es el siguiente:

$$Y = a + b * X + U$$

Donde

- Y es la variable respuesta o dependiente.
- X es la variable independiente o explicativa.
- U una variable aleatoria.

Es decir, el valor de la variable Y va depender del valor que toma la variable X , más una cierta perturbación aleatoria U, denominada residuo; gráficamente se puede representar de la siguiente manera:



Por lo tanto, los efectos que influyen en el valor que va a tomar la variable respuesta peso (Y) se pueden descomponer en dos factores o componentes:

- Efecto debido a la variable explicativa estatura (X), de la cual vamos a conocer su valor.
- Efecto producido por la perturbación aleatoria U, la cual recoge un conjunto muy grande de factores que influyen cada uno de ellos de forma moderada en la variable Y.

Las hipótesis consideradas en un modelo de regresión lineal simple son las siguientes:

a) **Linealidad**: vamos a considerar que los datos a analizar han sido generados por el siguiente modelo:
   - $Y = a + b * x + u$

b) **Homocedasticidad**: la variabilidad de la perturbación es constante y no depende de X.
   - $V(U) = \sigma^2$

c) **Independencia**: las perturbaciones aleatorias son independientes entre si.
   - $E(u_i \cdot u_j) = 0, \forall i \neq j$

d) **La perturbación aleatoria tiene de media 0**.
   - E(U)=0.

e) **Normalidad**: la distribución de la perturbación aleatoria tiene distribución normal.
   - $U \approx N(0, \sigma^2)$

III. COMENTARIOS A LA HIPÓTESIS

a) Linealidad

La hipótesis de linealidad establece que los todas las observaciones han sido generadas siguiendo un modelo lineal.
En la realidad será muy difícil aceptar esta hipótesis, pues raramente será creíble. En nuestro ejemplo, la relación del peso con la altura sabemos que no es lineal sino cúbica. Sin embargo, es fácil aceptar que, a grandes rasgos y en el intervalo muestreado, la relación puede interpretarse como lineal.

b) Homocedasticidad

La hipótesis de homocedasticidad establece que la variabilidad de los errores no va a depender de los valores de la variable independiente, aunque esto no siempre será cierto. En general, valores bajos de X implicará poca variabilidad de la perturbación y, análogamente, valores altos de X implicarán mayor variabilidad.

c) Independencia de las perturbaciones aleatorias

La hipótesis de independencia de las perturbaciones aleatorias establece que el valor de la perturbación aleatoria en el caso **i** no va a estar correlacionada con el valor de la perturbación en el caso **i+1, i+2, … i+k** . Cuando la variable tiempo está asociada de alguna manera con la variable independiente X esto no será cierto y deberemos de utilizar otros modelos estadísticos.

d) Residuos con media cero

Al considerar que los residuos tienen media cero para cualquier valor de X, el valor esperado de Y para un determinado valor de X va a ser:

$$E(Y/x) = E(a + b*X + U) = a + b*X + E(U) = a + b*X$$

$$E(Y/x) = a + b*X$$

Y continuando con nuestro ejemplo:

$$E(Peso/Estatura) = a + b*Estatura$$

e) Normalidad

Si la perturbación aleatoria tiene una distribución (aproximadamente) normal, será posible realizar inferencias sobre los parámetros de la recta de regresión, es decir, podremos asignar medidas de probabilidad a ciertas afirmaciones o hipótesis realizadas sobre los valores que pueden tomar los parámetros.

IV. ESTIMACIÓN DE LOS PARÁMETROS

La estimación de los parámetros de la recta de regresión puede realizarse por dos métodos:

- Mínimos cuadrados
- Máxima verosimilitud

Aunque estos métodos parten de hipótesis muy distintas, los estimadores de los parámetros van a coincidir.

IV.1. Mínimos cuadrados

En mínimos cuadrados, se estiman los parámetros de la recta de forma que se minimice la varianza de la perturbación aleatoria o varianza residual, es decir, se buscan unos valores **a** y **b** de forma que la distancia de cada punto a la recta de regresión (o valor pronosticado) sea mínimo; esto equivale a minimizar la varianza residual, teniendo en cuenta que la media de la perturbación es 0.

$$Min(\sigma^2) = Min\sum \frac{u^2}{n} = Min\sum \frac{(Y-(a+b*X))^2}{n}$$

Es decir, se calcularán los parámetros de la recta de regresión de forma que dicha recta pase lo más cerca posible (en promedio) de todos los puntos.

Derivando la expresión anterior respecto de los parámetros, igualándola a cero y aplicando un poco de álgebra obtenemos la expresión de los estimadores **a** y **b**.

Derivando respecto de **a** :

$$\frac{\partial U^2}{\partial a} = 2\sum (y - a - b*x) = 0$$

Se sigue la siguiente ecuación:

$$\boxed{\bar{y} = a + b*\bar{x}}$$

Derivando respecto de **b** :

$$\frac{\partial U^2}{\partial b} = 2\sum (y - a - b*x) \ *x = 0$$
$$\sum (y*x - a*x - b*x^2) = 0$$
$$\sum y*x = a\sum x + b\sum x^2$$
$$\sum \frac{y*x}{n} = a*\overline{X} + b\sum \frac{x^2}{n}$$

Multiplicando por $\overline{X}$ la primera ecuación y restándola a esta última obtenemos:

$$\text{cov}(x, y) = b*S_x^2 \Rightarrow$$

$$\boxed{b = \frac{\text{cov}(x, y)}{S_x^2}}$$

Luego, hemos conseguido poner los parámetros de la recta de regresión en función de las medias, varianzas y covarianzas muestrales.

Tabla resumen

b	$\frac{\text{cov}(x, y)}{S_x^2}$
a	$\overline{Y} - \frac{\text{cov}(x, y)}{S_x^2}\overline{X}$
$a + b*x$	$\overline{Y} - \frac{\text{cov}(x, y)}{S_x^2}\overline{X} + \frac{\text{cov}(x, y)}{S_x^2}*x$

Es importante observar que, en la estimación de los parámetros por mínimos cuadrados, no tenemos en cuenta las hipótesis anteriormente expuesta, en concreto la hipótesis de normalidad.

También es importante observar que el signo de la covarianza entre las dos variables va a determinar en gran medida el sentido de la recta de regresión.

IV.2. b Estimación por máxima verosimilitud

Si partimos de la base que la perturbación aleatoria sigue una distribución normal, podemos aplicar el método de estimación de máxima verosimilitud. La función de densidad para un caso concreto (o una persona concreta) es:

$$l(a,b,\sigma^2,y)=\frac{1}{\sigma\sqrt{2*\pi}}\exp\left(\frac{-1}{2*\sigma^2}(y-a-b*x)^2\right)$$

De donde la función soporte de la muestra, es decir, la verosimilitud de la muestra en estudio es:

$$l(a,b,\sigma^2,Y)=\frac{1}{\sigma^n\sqrt{(2*\pi)^n}}\prod\exp\left(\frac{-1}{2*\sigma^2}(y-a-b*x)^2\right)$$

El logaritmo de la función soporte será:

$$L(a,b,\sigma^2,Y)=\frac{-n}{2}\log(\sigma^2)-\frac{n}{2}\log(2*\pi)-\frac{1}{2\sigma^2}\sum(y-a-b*x)^2$$

Derivando esta expresión respecto de los parámetros **a** y **b**, e igualando a cero, de forma que se obtenga unos parámetros que maximicen la verosimilitud de la muestra, obtendremos:

$$\frac{\partial L}{\partial a}=2\sum(y-a-b*x)(-1)=0$$

$$\frac{\partial L}{\partial b}=2\sum(y-a-b*x)(-x)=0$$

Al coincidir estas ecuaciones con las obtenidas por mínimos cuadrados se sigue que los estimadores calculados por ambos métodos van a coincidir.

V. VARIANZA RESIDUAL

Vamos a descomponer la variabilidad de la variable dependiente peso (Y) en dos componentes, una componente explicada por el modelo de regresión y otra componente no explicada o aleatoria.

Empezamos por considerar la variabilidad de Y respecto a la media:

$$n * \sigma_y^2 = \sum (y_i - \overline{Y})^2$$

Es decir, la variabilidad de Y es la suma cuadrática de cada valor de y respecto a su media. Sumando y restando el valor pronosticado por la recta de regresión obtenemos la siguiente igualdad conocida como el teorema fundamental de la descomposición de la suma de cuadrados:

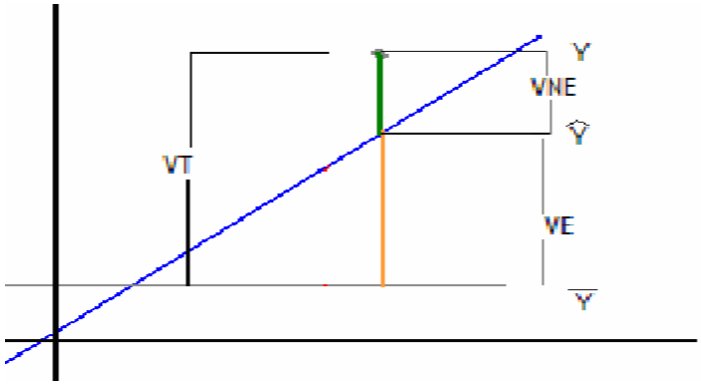
$$\sum (y_i - \overline{y})^2 = \sum (\hat{y}_i - \overline{y})^2 + \sum (y_i - \hat{y}_i)^2$$

Es decir, que la suma de cuadrados de la variable Y respecto a su media se puede descomponer en términos de la varianza residual. De esta expresión se deduce que “la distancia de Y a su media se descompone como la distancia de Y a su estimación más la distancia de su estimación a la media”.

Teniendo en cuenta que el último término representa la varianza no explicada, tenemos:

$$VT = VE + VNE$$

Gráficamente es fácil ver la relación:



Dividiendo la variabilidad total entre sus grados de libertad obtenemos la varianza estimada de la variable dependiente Y (Peso):

$$S_Y^2 = \frac{VT}{n-1}$$

Dividiendo la variabilidad no explicada entre sus grados de libertad obtenemos la varianza residual de la variable dependiente Y (Peso):

$$S_R^2 = \frac{VNE}{n-2}$$

Dividiendo la variabilidad explicada entre sus grados de libertad, obtenemos el estimador de la varianza explicada.

$$S_{VE}^2 = \frac{VE}{2-1}$$

Tabla resumen

Fuentes	Suma de cuadrados	Grados de libertad	Estimadores
VT	$\sum (y-\bar{y})^2$	n-1	$S_Y^2 = \frac{VT}{n-1}$
VE	$\sum (\hat{y}-\bar{y})^2$	2-1	$S_{VE}^2 = \frac{VE}{2-1}$
VNE	$\sum\ (y-\hat{y})^2$	n-2	$S_R^2 = \frac{VNE}{n-2}$

VI. CONTRASTE DE REGRESIÓN

Como estamos sacando conclusiones de una muestra de un conjunto mucho más amplio de datos, a veces este conjunto será infinito, es obvio que distintas muestras van a dar distintos valores de los parámetros.

Un caso de especial interés es asignar una medida de probabilidad a la siguiente afirmación o hipótesis:

$$H_0 \Rightarrow b = 0$$

Se denomina **contraste de regresión** al estudio de la posibilidad de que el modelo de regresión sea nulo, es decir, los valores de la variable Estatura (X) no van a influir en la variable Peso (Y). Teniendo en cuenta el modelo, esto es equivalente a la siguiente afirmación:

$$b = 0$$

Si esto es cierto, se sigue:

$$\hat{Y} = a + b * X$$
$$\hat{Y} = a + 0 * X$$
$$\hat{Y} = a \Rightarrow a = \overline{Y}$$

Es decir, el conocimiento de la estatura de una persona no va a proporcionar más información respecto de su peso que el conocimiento de la media de su peso.

Construcción del contraste

Si los residuos siguen una distribución normal y b=0, tenemos que:

$$\frac{VT}{\sigma^2} \approx \chi^2_{n-1}$$

$$\frac{VE}{\sigma^2} \approx \chi^2_{1}$$

$$\frac{VNE}{\sigma^2} \approx \chi^2_{n-2}$$

Por tanto:

$$\frac{VE/1}{VNE/n-2} = \frac{VE}{S_R^2} \approx F_{1,n-2}$$

Es decir, el cociente entre la varianza explicada y la varianza no explicada será aproximadamente 1. Además, al seguir una distribución F, podemos asignar una medida de probabilidad (p-value) a la hipótesis de que la varianza explicada es igual a la varianza no explicada.

En caso contrario la varianza no explicada será muy inferior a la varianza explicada y, por lo tanto, este cociente tendrá un valor muy superior a 1.

Nota

En general, si el p-value es menor de 0.05 se acepta que el modelo de regresión es significativo; en caso contrario no podemos hablar de regresión pues el modelo sería nulo.

Si aceptamos que el modelo de regresión es significativo, es habitual mostrar el p-value, por ejemplo:

Encontramos que este modelo de regresión es estadísticamente significativo con un p-value de 0.0003

VII. COEFICIENTE DE DETERMINACIÓN R^2

Vamos a construir un coeficiente (estadístico) que mida la bondad del ajuste del modelo. Si bien la varianza residual (S_R^2) nos indica cómo están de cerca las estimaciones respecto de los puntos, esta varianza está influida por la varianza de la variable dependiente, la cual, a su vez, está influida por su unidad de medida. Por lo tanto, una medida adecuada es la proporción de la varianza explicada (VE) entre la varianza total (VT); por ello, definimos el coeficiente de determinación R^2:

$$VT = VE + VNE$$

$$\frac{VT}{VT} = \frac{VE}{VT} + \frac{VNE}{VT}$$

$$1 = \frac{VE}{VT} + \frac{VNE}{VT}$$

$$\frac{VE}{VT} = 1 - \frac{VNE}{VT}$$

$$R^2 = \frac{VE}{VT} = \frac{VT - VNE}{VT} = 1 - \frac{VNE}{VT}$$

Por ser cociente de sumas de cuadrados, este coeficiente será siempre positivo.

Si todos los puntos están sobre la recta de regresión, la varianza no explicada será 0, y por lo tanto:

$$R^2 = \frac{VE}{VT} = 1 - \frac{0}{VT} = 1$$

Este coeficiente es muy importante, pues determina qué porcentaje (en tantos por uno) de la varianza de la variable dependiente es explicado por el modelo de regresión.

En general, se pueden clasificar los valores de R^2 de la siguiente manera:

Menor de 0.3	0.3 a 0.4	0.4 a 0.5	0.5 a 0.85	Mayor de 0.85
Muy malo	Malo	Regular	Bueno	Sospechoso

Además, a diferencia de la varianza residual, este coeficiente es **adimensional**, esto quiere decir que no se ve afectado por transformaciones lineales de las variables; por tanto, si cambiamos las unidades de medida, el coeficiente de determinación permanecerá invariante.

VIII. EJEMPLO 1

Los datos mostrados a continuación contienen las siguientes medidas de una muestra de 27 individuos.

Datos mostrados en parte.

estatura	sexo	peso	Pie	L_brazo	a_espalda	d_craneo	L_roxto
159	0	49	36	68.5	42	57	40
164	1	62	39	73	44	55	44
172	0	65	38	75	48	58	44
167	0	52	37	73	41.5	58	44

1. ESTATURA — Estatura en centimetros
2. SEXO — Sexo del individuo
   - 0 Mujer
   - 1 Hombre
3. PESO — Peso en kilogramos
4. PIE — Talla del pie
5. L_BRAZO — Longitud del brazo
6. A_ESPALD — Ancho de la espalda
7. D_CRÁNEO — Diametro del craneo
8. L_ROXTO — Longitud de rodilla a tobillo

Deseamos realizamos una regresión de la estatura sobre el peso, basado en una muestra de 27 personas de ambos sexos.

El peso va medido en kilogramos y la altura en centímetros.

Resultados de la regresión

a) Coeficiente de determinación

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,832[a]	,692	,680	7,23925

a. Predictors: (Constant), estatura

El coeficiente de determinación (R Square) es de 0.692. Por lo tanto el 69% de la varianza del peso de una persona queda explicado por su estatura.

La desviación típica residual (S_R) es 7.239.

b) Contraste de regresión.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2950,497	1	2950,497	56,300	,000[a]
	Residual	1310,170	25	52,407		
	Total	4260,667	26			

a. Predictors: (Constant), estatura

b. Dependent Variable: peso

Aquí tenemos las sumas de cuadrados en el siguiente orden:

VE	Regresión
VNE	Residual
VT	Total

Fijémonos que VT = VNE+VE.

Dividiendo VNE entre sus grados de libertad, obtenemos la varianza residual.

$$S_R^2 = \frac{VNE}{n-2} = \frac{\text{Re}\,sidual}{27-2} = \frac{1310.17}{25} = 52.407$$

Y la raíz cuadrada de la varianza residual es el error típico de la estimación:

$$Sd\,\text{Re}\,sidual = \sqrt{52.407} = 7.24$$

Dividiendo la varianza explicada entre la varianza no explicada obtenemos el contraste de regresión:

$$F = \frac{VE}{S_R^2} = \frac{2950.497}{52.407} = 56.3$$

Buscando en la tabla de distribución acumulada de una distribución F con 1 y 25 grados de libertad el valor 56.3 obtenemos el p-value:

$$P(F > 56.3) = 7{,}4072084\,8253e\text{-}008$$

Por lo tanto se concluye que es muy poco probable que estos datos hayan sido extraídos de una población donde b=0.

Dividiendo la variabilidad explicada entre la variabilidad total obtenemos el coeficiente de determinación.

$$R^2 = \frac{VE}{VT} = \frac{\text{Re}\,gresión}{Total} = \frac{2950}{4260} = 0.692$$

c) Estimación de los coeficientes.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-112,039	23,488		-4,770	,000
	estatura	1,042	,139	,832	7,503	,000

a. Dependent Variable: peso

El modelo de regresión estimado es el siguiente:

$$Peso = -112.04 + 1.042 * estatura$$

Se interpreta que por cada centímetro que incrementa una persona su estatura va a aumentar en 1.042 kilogramos su peso. Por ejemplo, el peso estimado para una persona de 176 cm. de estatura será:

$$\text{E(Peso/estatura = 176)} = -112.04 + 1.042 * 176 = 71.352$$

El intervalo de confianza al 95% de esta estimación será:

$$\hat{Y} \pm 1.96 * \sqrt{S_R^2} = 1.96 * 7.2 = 14.112$$

$$71.352 \pm 14.112$$

El **error estándar** de los estimadores de los coeficientes se muestra en la segunda columna.

La tercera columna es el **coeficiente b estandarizado**, esto es equivalente al coeficiente que se obtendría al realizar el modelo de regresión con las variables estandarizadas. Su interés radica en poder comparar el impacto de varias variables explicativas independientemente de su unidad de medida.

La cuarta columna es el **valor de los estimadores** dividido entre los estimadores de su desviación típica; este cociente se suele denominar con la letra **T** ;el interés de este cociente es contrastar la hipótesis de que dicho coeficiente es nulo.

IX. EJEMPLO 2: TEMPERATURA DE EBULLICIÓN

Los siguientes datos fueron tomados por el físico escocés Forbes en los Alpes en 1857, que midió la temperatura de ebullición del agua en grados Fahrenheit y la presión atmosférica.

Datos:

PRESION	20.79	22.40	23.15	23.89	24.02	25.14	28.49	29.04	29.88	30.06
TEMP	194.50	197.90	199.40	200.90	201.40	203.60	209.50	210.70	211.90	212.20

Se desea estudiar la relación funcional temp =f (presión) mediante un modelo de regresión lineal.

Generalmente, las fases de un análisis de regresión lineal simple son las siguientes:

1. Solicitar un análisis descriptivo de las variables involucradas en el modelo; los estadísticos solicitados deberán de ser los siguientes: mínimo, máximo, media, mediana, simetría y curtosis.

2. Realizar un gráfico de dispersión, poniendo la variable explicativa en el eje horizontal y la dependiente en el eje vertical.

3. Solicitar las correlaciones entre las dos variables.

4. Realizar el análisis de regresión lineal simple solicitando los siguientes subanálisis:
   - Coeficiente de correlación de los residuos Durbin-Watson
   - Histograma de los residuos
   - Gráfico de valores pronosticados VS residuos al cuadrado

5. Si se observan casos que tienen gran influencia en el modelo de regresión, se deberá de plantear su posible exclusión del mismo.

6. Interpretar los coeficientes del modelo mostrando las medidas de bondad del ajuste y el contraste de regresión.

Resultados del análisis:

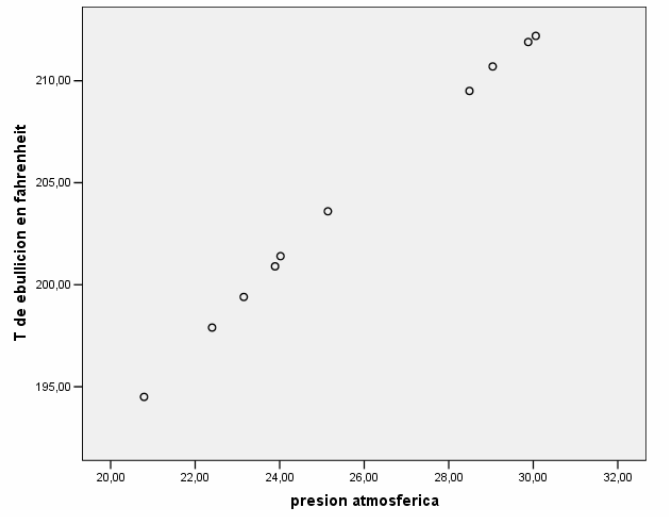
a) Análisis descriptivo

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
presion presion atmosferica	10	20,79	30,06	25,6860	3,38842	,134	,687	-1,666	1,334
tem T de ebullicion en fahrenheit	10	194,50	212,20	204,2000	6,40781	,032	,687	-1,579	1,334
Valid N (listwise)	10								

Aparentemente los valores están en rango y las distribuciones no son asimétricas, así mismo no se detectan valores altos de curtosis.

b) Gráfico de dispersión



La relación entre presión y temperatura es aparentemente lineal, no se detectan casos atípicos que se salgan fuera de la norma general.

c) Matriz de correlaciones

Correlations

		presion presion atmosferica	tem T de ebullicion en fahrenheit
presion presion atmosferica	Pearson Correlation	1	,999**
	Sig. (2-tailed)		,000
	N	10	10
tem T de ebullicion en fahrenheit	Pearson Correlation	,999**	1
	Sig. (2-tailed)	,000	
	N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

La correlación entre las variables es altísima 0.999.

d) Análisis de regresión propiamente dicho

- <u>Bondad del ajuste</u>

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,999[a]	,998	,998	,29005	,860

a. Predictors: (Constant), presion presion atmosferica

b. Dependent Variable: tem T de ebullicion en fahrenheit

Como el coeficiente de determinación es 0.998 quiere decir que con el conocimiento de la presión atmosférica podemos explicar el 99.8% de la variabilidad de la temperatura de ebullición.

En cambio DW tiene un valor muy alejado de 2, indicando que los residuos están fuertemente correlacionados, en este caso es debido a que la relación no es del todo lineal.

- <u>Contraste de regresión</u>

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	368,867	1	368,867	4384,568	,000[a]
	Residual	,673	8	,084		
	Total	369,540	9			

a. Predictors: (Constant), presion presion atmosferica

b. Dependent Variable: tem T de ebullicion en fahrenheit

El contraste de regresión es significativo, indicando por tanto que el modelo es estadísticamente significativo con un p-value de 4.1E-29.

- Coeficientes estimados

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	155,670	,739		210,756	,000
	presion presion atmosferica	1,889	,029	,999	66,216	,000

a. Dependent Variable: tem T de ebullicion en fahrenheit

El modelo de regresión estimado es:

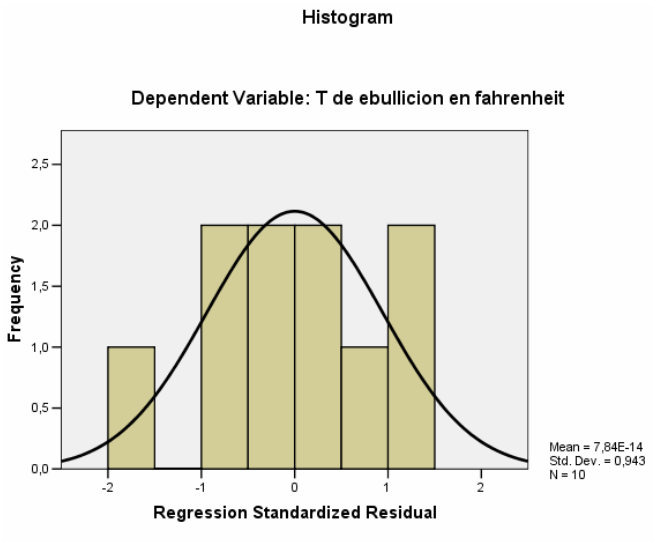
$$Temp = 155.7 + 1.9 * presion$$

Para una presión de cero, la temperatura de ebullición deberá de ser de 155.7

Residuals Statistics[a]

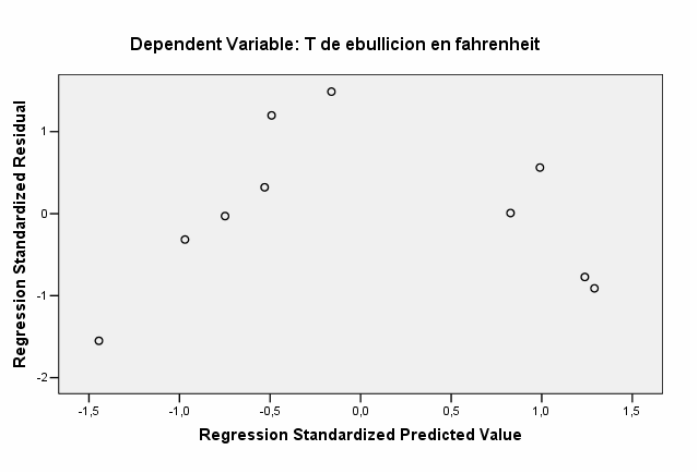
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	194,9497	212,4641	204,2000	6,40197	10
Std. Predicted Value	-1,445	1,291	,000	1,000	10
Standard Error of Predicted Value	,093	,167	,128	,024	10
Adjusted Predicted Value	195,1731	212,5694	204,2279	6,39021	10
Residual	-,44965	,43159	,00000	,27346	10
Std. Residual	-1,550	1,488	,000	,943	10
Stud. Residual	-1,897	1,571	-,042	1,071	10
Deleted Residual	-,67311	,48109	-,02793	,35487	10
Stud. Deleted Residual	-2,392	1,767	-,068	1,210	10
Mahal. Distance	,026	2,088	,900	,679	10
Cook's Distance	,000	,894	,162	,269	10
Centered Leverage Value	,003	,232	,100	,075	10

a. Dependent Variable: tem T de ebullicion en fahrenheit

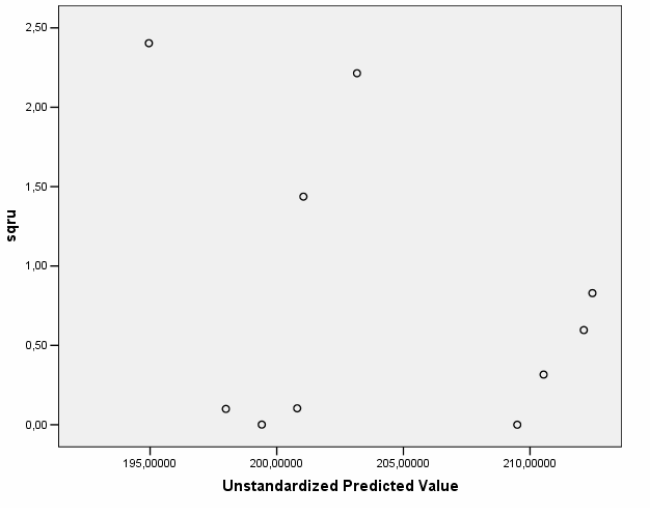


En el histograma de los residuos se observa que si bien no son normales tampoco se desvían excesivamente de la normal; en cualquier caso, no hay suficientes casos como para afirmar o negar esta hipótesis de forma categórica.

La falta de normalidad puede estar afectada por la falta de linealidad.



En el gráfico de valores pronosticados vs residuos queda clara la falta de linealidad.



En el gráfico de valores pronosticados vs residuos al cuadrado no se aprecia falta de homocedasticidad.